AD/A-003 034

INTERACTIVE SYSTEMS RESEARCH

M. I. Eernstein

System Development Corporaιion

Prepared for:

Advanced Research Projects Agency

15 November 1974

DISTRIBUTED BY:

NTIS

National Technical Information Service
U. S. DEPARTMENT OF COMMERCE

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER | 2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER AD/A-003034
4. TITLE (and Subtitle) Interactive Systems Research: Final Report to the Director, Advanced Research Projects Agency, for the Period 1 October 1973 to 15 September 1974	5. TYPE OF REPORT & PERIOD COVERED Final--10/1/73 - 9/15/74
	6. PERFORMING ORG. REPORT NUMBER TM-5243/002/00
7. AUTHOR(s) Bernstein, M. I.	8. CONTRACT OR GRANT NUMBER(s) DAHC15-73-C-0080
9. PERFORMING ORGANIZATION NAME AND ADDRESS System Development Corporation 2500 Colorado Avenue Santa Monica, California 90406	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 5D30
11. CONTROLLING OFFICE NAME AND ADDRESS Information Processing Techniques Office Advanced Research Projects Agency 1400 Wilson Blvd., Arlington, Virginia	12. REPORT DATE 15 November 1974
	13. NUMBER OF PAGES 43
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report) Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Reproduced by
NATIONAL TECHNICAL
INFORMATION SERVICE
US Department of Commerce
Springfield, VA 22151

TM-5243/002/00

INTERACTIVE SYSTEMS RESEARCH: FINAL REPORT TO THE DIRECTOR, ADVANCED RESEARCH PROJECTS AGENCY, FOR THE PERIOD 1 OCTOBER 1973 to 15 SEPTEMBER 1974

15 NOVEMBER 1974

M. I BERNSTEIN
(213) 393-9411, EXT. 6117

THIS REPORT WAS PRODUCED BY SDC IN PERFORMANCE OF CONTRACT DAHC15 73-C-0080. ARPA ORDER NO. 2254 PROGRAM CODE NUMBER 5D30

THE VIEWS AND CONCLUSIONS CONTAINED HEREIN ARE THOSE OF THE AUTHOR AND SHOULD NOT BE INTERPRETED AS NECESSARILY REPRESENTING THE OFFICIAL POLICIES EITHER EXPRESSED OR IMPLIED OF THE ADVANCED RESEARCH PROJECTS AGENCY OR THE U S GOVERNMENT.

System Development Corporation
2500 Colorado Avenue • Santa Monica, California 90406

TABLE OF CONTENTS

1. INTRODUCTION AND SUMMARY 3

1.1 Speech Understanding Research 4
1.2 Lexical Data Archive 4
1.3 Common Information Structures 4

2. SPEECH UNDERSTANDING RESEARCH 5

2.1 Introduction 5
2.2 Progress for the First Six-Month Period 5

2.2.1 System Overview 6
2.2.2 The Discourse-Level Controller 8
2.2.3 The Parser 9
2.2.4 The Lexical Matching Procedure 11
2.2.5 The Acoustic-Phonetic Processor 12
2.2.6 System Testing 15
2.2.7 Related Research 16

2.3 Progress for the Final Six-Month Period 17

2.3.1 Demonstration System Overview 17
2.3.2 Acoustic-Phonetic Research 18
2.3.3 System Hardware and Software 20

2.4 Plans 21
2.5 Staff 23

3. LEXICAL DATA ARCHIVE 24

3.1 Introduction 24
3.2 Progress and Present Status 24

3.2.1 Completion of SOLAR Design 25
3.2.2 Hand Collection of Data 25
3.2.3 Machine Derivation of Data 27
3.2.4 Programs under Development 27
3.2.5 Data Management 27
3.2.6 Data Distribution 28

3.3 Plans 28
3.4 Staff 29

4. COMMON INFORMATION STRUCTURES 30

4.1 Introduction 30
4.2 Progress and Present Status 31
4.3 Plans 35

4.3.1 Implementation of the Logical Data Translator 35
4.3.2 Design of the Reformatters 36
4.3.3 Experimental Conversion of a Data Base 36

4.4 Staff 36

5. REFERENCES 37

APPENDIX 40

1. INTRODUCTION AND SUMMARY

This report to the Advanced Research Projects Agency (ARPA) is the final one for the period 1 October 1973 through 15 September 1974 on System Development Corporation's (SDC's) research and development program in Interactive Systems Research (ISR). An interim report covering the first six months of this program (through 31 March 1974) has been submitted (Bernstein, 1974). The present report concentrates on the progress, results, and problems for the last half of the year's work and contains plans for the future activities of the ISR program. The program presently includes three projects: (1) Speech Understanding Research, (2) Lexical Data Archive, and (3) Common Information Structures. The overall intent of SDC's ISR program is to develop basic technology for improved man-machine interactive systems with application to a variety of anticipated military needs. The major emphasis at present is on speech understanding and related problems.

The Speech Understanding Research project contains many developments that will be material in enhancing and improving interactive systems by making them more capable and productive when used by the casual user. The continuing effort to permit such users to easily and effectively communicate with a computer-based system in language forms that are natural to them is of particular importance.

In support of ARPA's Speech Understanding Research program and other language-based efforts, the Lexical Data Archive project was started to create (as its name implies) a central archive of lexical information for, and of particular interest to, all ARPA contractors working in speech understanding as well as other language-based activities.

The world of information processing has been one of continuous evolution and change since its creation. The ever increasing dependence of users of information processing systems upon data bases and data management systems has become obvious. With the continuously changing environment of hardware, operating systems, and data management systems within which data bases reside, the need to move a data base with minimal effort, cost, and disruption to its users is becoming as important as smooth, well-engineered user interfaces. The Common Information Structures project is continuing to develop the methodology necessary to perform data base transfers in the appropriate way.

The following summarizes these projects' activities during the past year.

1.1 SPEECH UNDERSTANDING RESEARCH

The Speech Understanding Research (SUR) project is continuing its efforts to create a demonstrable prototype Voice-controlled Data Management System (VDMS), using free-form spoken English as input. Although it was originally intended that an enhanced version of the prototype that was demonstrated to the SUR Group Review Team in December, 1973, would be demonstrable in the fall of 1974, the effort has been redirected to produce a version of VDMS later in 1974 that incorporates the linguistic processor being developed by the Stanford Research Institute (SRI) in conjunction with SDC's SUR project, along with improvements that have been made in SDC's phonological and acoustic-phonetic processes. The development work is all being done on the SDC Computer Center Facility's IBM 370/145 VM-370 Operating System with remote users accessing the system via the ARPA Network.

1.2 LEXICAL DATA ARCHIVE

The Lexical Data Archive (LDA) project was begun in October, 1973, with the objective of providing, via the ARPA Network, a centrally available collec ion of lexical information on the union of lexicons used by the ARPA SUR contractors in their collective and individual endeavors. The Semantically Oriented Lexical Archive (SOLAR) has been designed, and most of the relevant data have been collected. Programs have been developed for deriving machine-readable data files and for discovering and displaying links among lexicon words, and a file management facility has been developed. To date, data have been manually distributed to the SUR projects and others. Automatic access will be provided in the near future.

1.3 COMMON INFORMATION STRUCTURES

The COmmon INformation Structures (COINS) project has devised a method for transferring data bases between disparate data management systems (DMSs) that requires minimal effort and cost by utilizing to a maximum degree the functional capabilities of the two DMSs involved. The method depends upon the existence of three languages to describe the various phases of the intermediate process of the actual transfer. They are a Common Data Description Language, a Common Data Translation Language, and a Common Data Format Language. These languages have been defined, and effort is now concentrating on refinement and the implementation of language processors that will permit a thorough test and evaluation of the method.

2. SPEECH UNDERSTANDING RESEARCH

2.1 INTRODUCTION

The continuing long-term goal of the SDC Speech Understanding Research (SUR) project is to develop and implement a data management system that is controlled and operated by its users through free-form spoken English. The basic approach taken to achieve this goal is distinguished by a modular system architecture that embodies phonological and linguistic processes and an acoustic-phonetic processor. The system architecture enables a complete assembly of multidirectional parsing processes to operate in parallel on the same or different segments of an input utterance. Two major advantages are obtained from this: (1) the system may start working on the least ambiguous portions of the input, and (2) predictions need not be limited to near neighbors of a recognized input segment but may be applied to any portion of the entire utterance.

The acoustic-phonetic processor contains the processes that extract acoustic information from the speech signal and make acoustic-phonetic labeling decisions. The processor we are developing reflects the fact that the speech signal is never wholly unambiguous: any attempt to precisely label phones and their boundaries must recognize and allow for this ambiguity in mapping the extremely large number of speech sounds into the relatively small set of acoustic-phonetic transciption symbols. Accordingly, in this processor, each acoustic-phonetic segment is multiply labeled, and each label is assigned a score. Scores are based on a measure function that is, in turn, based on feature parameters previously developed for each speaker (user).

As a first step toward the long-term goal, we constructed and refined a limited Voice-controlled Data Management System (VDMS) that could accept continuous speech and be demonstrably usable by at least two speakers. Within this system, there were limitations with respect to both the size of the vocabulary and the syntax of the English subset permitted.

2.2 PROGRESS FOR THE FIRST SIX-MONTH PERIOD

At the beginning of the present contract year, two separate versions of VDMS had been constructed and tested: Version A, which operated on the SUR laboratory Raytheon 704 minicomputer in conjunction with an IBM 370/145, and which incorporated the modular system architecture, and Version B, which operated entirely on the Raytheon 704, and which embodied the multiple-labeling philosophy of the acoustic-phonetic processor described above. Both versions allowed the user to access a data base of information about the length, beam, draft, armament, and other characteristics of submarines in the naval fleets of the United States, the Soviet Union, and the United Kingdom. The total vocabulary of each system was approximately 150 words. The query language used to access this information could be described by about 35 syntax equations for Version A and about 30 syntax equations for Version B; the difference is accounted for by the fact that Version A contained report generation capabilities that Version B did not. Typical queries that could be accommodated by either system are:

> "TOTAL QUANTITY WHERE TYPE EQUALS NUCLEAR AND COUNTRY EQUALS USA."
>
> "PRINT TYPE WHERE MISSILES GREATER THAN SEVEN."

Both of these initial versions of VDMS had been tested with a large number of utterances and had achieved reasonably good results (Bernstein, 1974). The first major task undertaken during this contract year was the construction of a single full-scale version of VDMS that combines the best elements of the two versions. This new version of VDMS (Ritea, 1974a,b) was completed and successfully demonstrated in late 1973. Its major characteristics are described in this section.

2.2.1 System Overview

The overall configuration of VDMS is characterized by three major processing modules:

(1) The linguistic processor, which contains the parser and a discourse-level controller;

(2) The acoustic-phonetic processor, whose results are contained in an array of data called the A-matrix;

(3) The lexical matching procedure, which performs matches of predicted words at the syllable level, using various applications of phonological rules to assist in its matchings.

The pattern of communication among these modules is illustrated in Figure 2-1. The speech from the user is input to the acoustic-phonetic processor, which forms an array of acoustic-phonetic data for use by the parser. At the beginning of the processing of an utterance, the discourse-level controller provides a variety of predictions and restrictions on what is allowable or expected in this utterance. The predicted words ar transmitted to the lexical matching procedure, which looks for

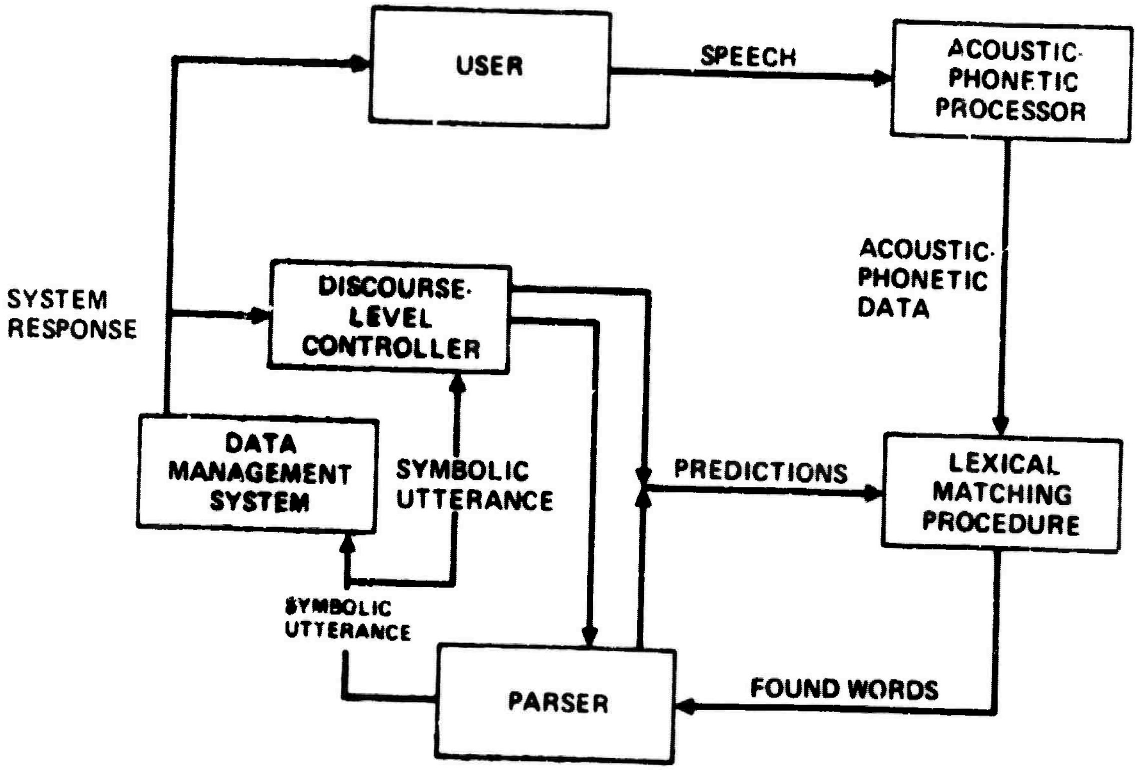

Figure 2-1. Overview of VDMS

the words in the acoustic-phonetic data. The parser and lexical matching procedure then pass predictions and verifications back and forth to one another in an effort to understand the utterance. Once it is understood, the utterance is passed to the data management system, which forms an appropriate response. The response is passed to the discourse-level controller and to the user. The discourse-level controller is then updated to aid in future predictions.

The logical flow of control and data between all of the modules is specified by a language unique to VDMS, called the Control Structure Language (CSL). Using CSL for program control, new modules may be implemented, and data paths among modules in VDMS may be modified without major reprogramming of the system. In addition, CSL has the following features:

(1) It allows for logical parallel execution of modules.

(2) It provides for the running of modules on remote computers.

(3) Using a trace and debugging provision, breakpoints can be inserted for monitoring the flow of data through the system.

(4) Changes in the order of execution of the various modules may be specified.

(5) Data dependencies among modules may be controlled.

A more detailed discussion of CSL is given by Barnett (1974b).

2.2.2 The Discourse-Level Controller

The discourse-level controller comprises two modules: the user model and the thematic memory. The user model determines what query state the user is in and predicts the kinds of grammar that may appear in his next interaction with the system. Some sample states are "system login", "interactive query mode", "report generation mode", and "user aids". If the user is in interactive query mode, the user model will predict syntax equations for the next interaction, such as those for "Print", "Repeat", "Count", "Subset", or "Total", each of which is the first word of an interactive query statement. The words "Explain" and "Describe" are the first words of typical user-aid commands. Each prediction carries with it a confidence level such that the higher the confidence level, the more liberal the system will be in overlooking errors in recognition.

The thematic memory is concerned with particular content words that might occur in the next utterance; it is not concerned with syntactic terminals such as the digits or the word "Print". Several pieces of information are kept about each word as it is used: e.g., how long (how many utterances ago) it has been since the word was used and how likely it is that the word will re-occur, depending on how it was used originally. For example, if the user said "show category", the assumption is that the next command will probably involve something about the categories of submarines in the data base.

In addition to looking for content words in user commands, the thematic memory also keeps a record of any non-numeric symbolic responses from the data management system. These responses are also used to predict words that are highly likely to occur in the next utterance.

Throughout a dialogue, the various content words as predicted by the thematic memory are aged from utterance to utterance, and their likelihood of being used is diminished if they have not been reused. If the confidence of prediction drops below a threshold, then the word is removed from the thematic memory and dropped from consideration until it is used again. Also, if duplicate entries occur within an utterance, then the age and original merit are modified to take care of this effect.

2.2.3 The Parser

The basic linguistic unit used for the parsing strategy in VDMS is the phrase, which consists of one or more vocabulary words (up to the complete utterance) linked together in a syntactically and semantically correct order. Some examples of phrases are "country and category" and "quantity equals five". The parser attempts to predict phrases using the user model, thematic patterning, and grammatical and semantic constraints information provided by the discourse-level controller. Predicted phrases are matched against the acoustic-phonetic data for acceptance or rejection. Accepted phrases are then concatenated to form a larger phrase, which is then analyzed to see whether it is a complete utterance.

The parser consists of four major modules:

(1) The classifier

(2) The bottom driver

(3) The top driver

(4) The side driver

The classifier's task is to assign a syntactic category to each word accepted by the lexical matching procedure. Some typical syntactic categories and examples are:

Item name ("country")

Item value ("USA")

Syntactic terminal ("print")

A syntactic category, as generated by the classifier, is used by the other modules of the parser to generate predictions about allowable syntax in other parts of the utterance. The bottom driver is a typical bottom-up module, which takes found phrases and determines how they may be used in completing the parsing of a complete utterance. The top driver takes predicted phrases and from them derives either a syntactic terminal or a shorter phrase to be looked for next.

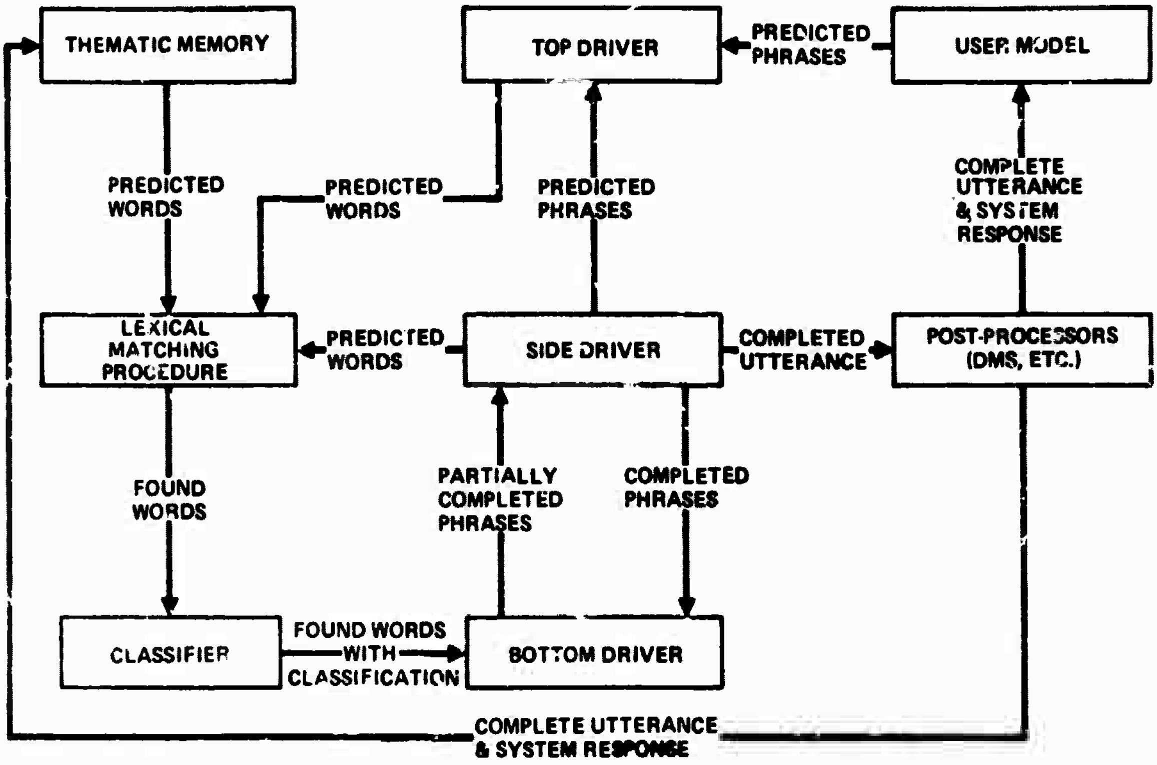

Figure 2-2. The Parser

The syntactic terminals are sent to the lexical matching procedure, which then attempts to match each one against the acoustic-phonetic data. The side driver takes completed or partially completed phrases from the bottom driver. If a phrase is incomplete, the side driver determines which part to look for next, and it will ask the top driver to locate the missing part. On the other hand, if the phrase has been completed, the side driver analyzes it to see whether it is a legal complete utterance. If it is, the side driver terminates the parsing activities of all modules and transmits the symbolic form of the hypothesized utterance to the data management system and the discourse-level controller. Other completed phrases (which do not cover the entire utterance) are used to bottom drive the system up one level to create larger, more complete phrases. The flow of processing within the parser is shown in Figure 2-2.

2.2.4 The Lexical Matching Procedure

The lexical matcning procedure verifies or rejects a predicted word through pattern-matching against the available acoustic-phonetic data. A detailed description of the procedure is given by Weeks (1974a,b).

The syllable is the unit that is used in the lexical matching process. The linguistic issues concerning the existence or form of the syllable have been sidestepped by giving it the following algorithmic definition: a vowel nucleus preceded by a consonant cluster (possibly null) and followed by another consonant cluster (also possibly null). All words have syllable divisions marked in the lexicon, and some of the phonemic rules are written in terms of these boundaries. Within a syllable, most of the co-articulation is internal;

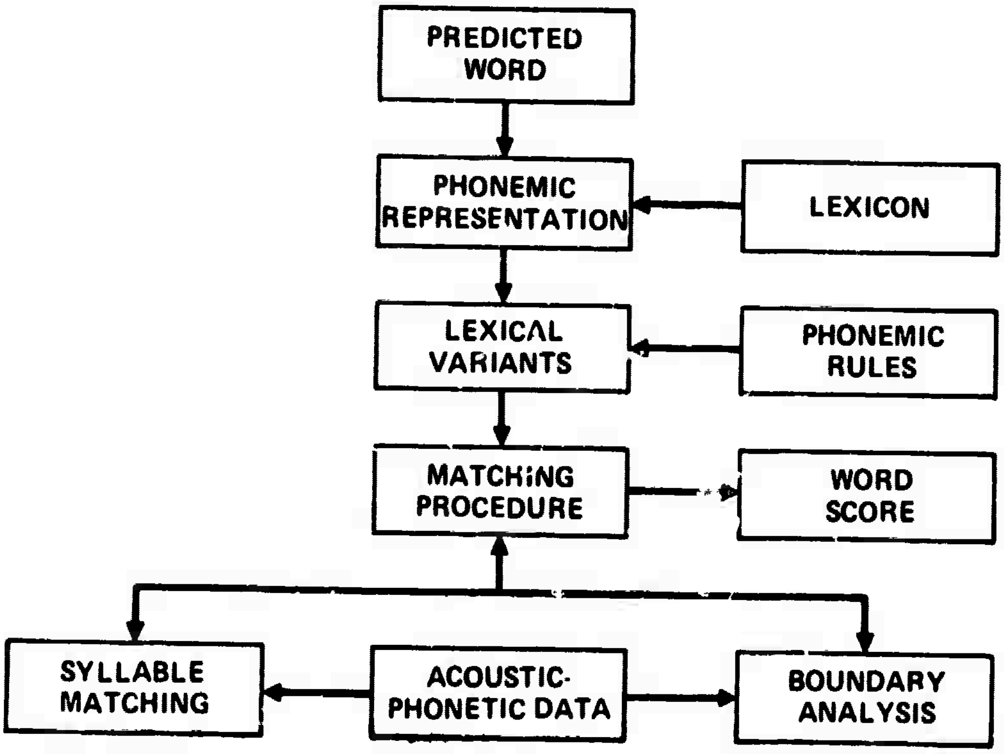

Figure 2-3. Lexical Matching Procedure

effects over boundaries are handled separately. Because a good percentage of phonetic dependencies occur within these units, rules can conveniently be applied. Under this approach, a separate set of rules must be set up for dealing with interactions over boundaries. Word boundaries can then be considered as special cases of syllable boundaries, so that the rules apply to inter-word co-articulation.

Figure 2-3 is a block diagram of the lexical matching procedure. When a word is predicted in orthographic form, its phonemic representation is extracted from the lexicon. A set of phonemic rules is applied to the phonemic representation to obtain a set of lexical variants. These variants arise by phonemic replacements as dictated by the rules. The set of lexical variants is then sent to the main matching procedure, where each is matched one by one against the acoustic-phonetic data on a syllable-by-syllable basis. The boundary analysis is done in conjunction with the syllable matching and attempts to compensate for articulation across syllable and word boundaries. The resulting scores for each lexical variant are then sent to the main matching procedure, which decides which possibility gives the best over-all score. The structure of the phonemic rules pass is described by Barnett (1974a).

2.2.5 The Acoustic-Phonetic Processor

Speech is input interactively in an acoustically controlled environment (a sound booth with a signal-to-noise ratio better than 50 dB) using a Sony ECM-377 condenser microphone, which has an essentially flat frequency response to beyond 10,000 Hz. Low-level preemphasis is employed, shaping the frequency response with a zero at 300 Hz and a pole at 3,000 Hz. This speech signal is bandlimited to 9,000 Hz and digitized at 20,000 samples per second using a 12-bit analog-to-digital converter. The input speech is saved directly on digital media with no intervening analog recording steps.

The digitized speech is then passed through a digital-to-analog converter, and the resulting analog waveform is passed through three hardware filters having nominal bandpasses of 150 to 900 Hz, 900 to 2,200 Hz, and 2,200 Hz to 5,000 Hz, respectively. For each 10-msec. interval, two parameters are extracted from each of the three filter outputs: (1) the maximum peak-to-peak amplitude and (2) a count of zero crossings. The resulting six parameters (two from each of the three filtered signals) are used to assign a rough acoustic label to each 10-msec. segment. Five labels are currently used: VW (vowel-like), SS (strong frication), SI (silence), UV (low-amplitude voiced or unvoiced), and VC (all other--usually weak voicing). The next step in the processing is to refine these rough labels, imposing a more accurate classification on each segment.

Within the classes VW and VC, more specific labels are assigned using a vowel-recognition strategy based on speaker-dependent vowel formant information (Kameny and Weeks, 1974). This information is compared with the formants (obtained with the use of a Linear Predictive Coefficient (LPC) spectrum) at selected instants in time for each vowel to be identified. A modified Euclidean distance function is used to compute the relative distances between the candidate formant values and pre-stored speaker-dependent vowel formant values. The closest three vowels are selected, and associated scores are assigned to these choices based on the values of the distance function.

Fricatives and plosives are characteristically found within sequences of segments labeled SS, VC, or UV. For these areas, a technique called the Low-Coefficient LPC (LCLPC), described by Molho (1974a,b), has been shown to provide meaningful spectra that correspond well with both acoustic-phonetic theory and with the experimental results of others. Time resolution is sufficiently narrow to allow independent spectral analysis of the release, frication, and aspiration portions of an unvoiced plosive or to demonstrate spectral change within a consonant cluster, so that clusters such as /ks/ and /ts/ may often be distinguished. For analysis of unvoiced speech, the LCLPC uses the autocorrelation method with eight coefficients and a 6-msec. Hamming window. Analysis of spectra obtained in this way allows the following five classes to be distinguished:

(1) labial or dental (LD)

(2) alveolar (AL)

(3) alveopalatal (AP)

(4) palatal or velar (PV)

(5) voiced or low energy (VS)

These classes correspond roughly to the spectral characteristics of unvoiced fricatives and plosives. Moreover, there is a correspondence between these classes and the articulatory positions of unvoiced fricatives and plosives. The classes LD, AL, AP, and PV ideally contain the following phonemes:

LD: /p/,/f/, /θ/

AL: /t/, /s/

AP: /ʃ/

PV: /k/

Experimentation has also confirmed that the glide /w/ and the liquid /l/ characteristically occur within the VW or VC classes. The present approach to recognizing these phonemes is to augment a speaker's vowel formant table with the formant frequency values for /w/ and /l/. These formant values have consistently been easily distinguishable from the formants of the vowels and have enabled the system to accurately isolate and recognize /w/ and /l/. The glide /y/ and the liquid /r/ are handled indirectly, again with the use of the speaker-dependent vowel formant table: if a 10-msec. segment has been labeled /i/, it is assumed that the segment could be a /y/ with equal probability, and both labels are then assigned to the segment with the same score. If a segment has been labeled /ɤ/ (again with the aid of the vowel table), the label /r/ is assigned to the same segment with an equal score.

Although the system is not yet able to distinguish the various elements within the class of nasals, viz., /m/, /n/, /ŋ/, /m̩/, /n̩/, a single class name (NA) is used and has proved quite reliable. A segment is labeled NA based upon some simple tests involving the amplitudes and bandwidths of formants F1, F2, and F3.

All of the aforementioned segment labeling procedures are used to construct an array of acoustic-phonetic data called the A-matrix. The construction of the A-matrix is shown in Figure 2-4. Each row of the A-matrix corresponds to a 10-msec. segment of speech and contains a rough segment label (VW, SS, SI, VC, or UV); one or more refined segment labels and associated scores based on the above procedures; formant frequency values; and estimates of fundamental frequency, RMS energy, and other acoustic-phonetic parameters used in the assignment of the phoneme and phoneme-class labels.

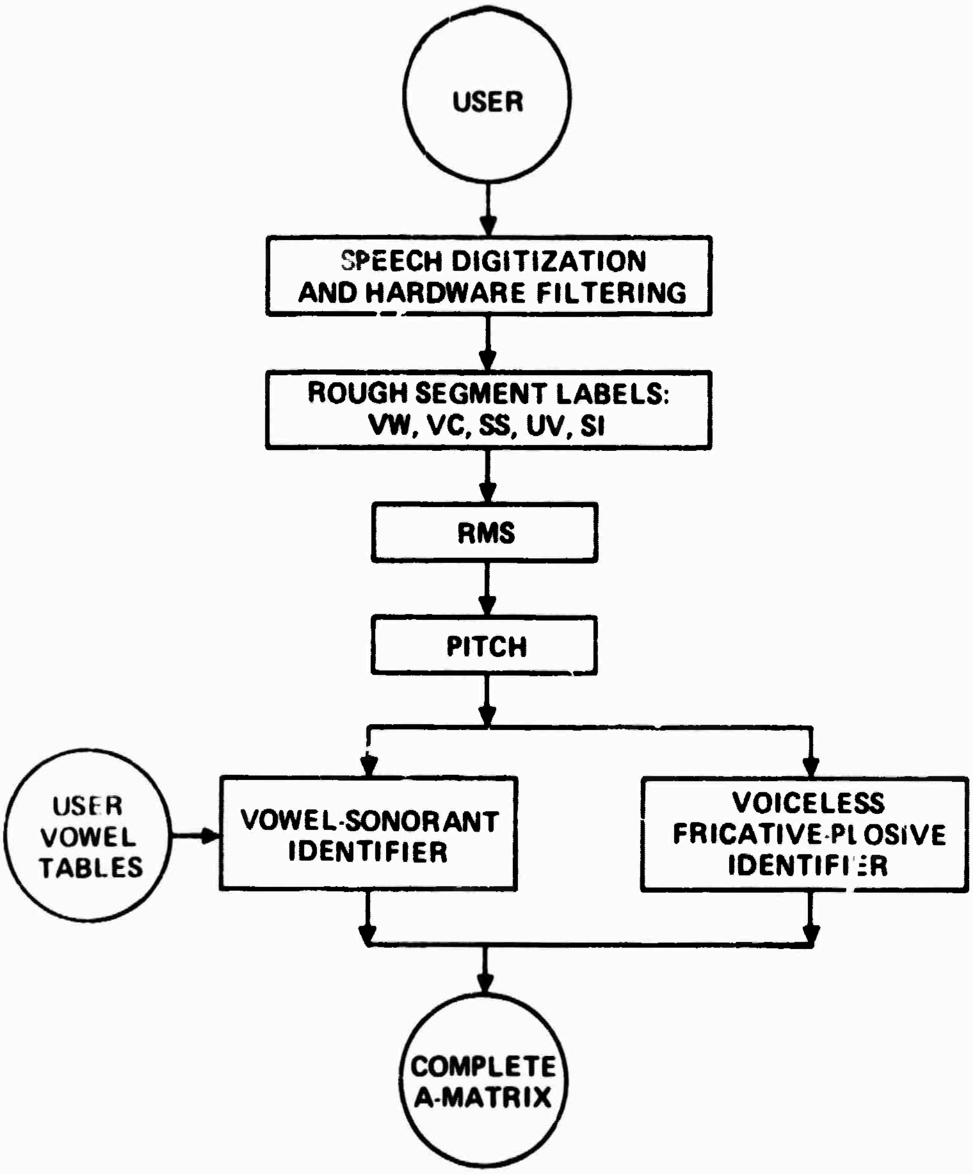
USER
SPEECH DIGITIZATION AND HARDWARE FILTERING
ROUGH SEGMENT LABELS: VW, VC, SS, UV, SI
RMS
PITCH
USER VOWEL TABLES
VOWEL-SONORANT IDENTIFIER
VOICELESS FRICATIVE-PLOSIVE IDENTIFIER
COMPLETE A-MATRIX

Figure 2-4. A-Matrix Processing

2.2.6 System Testing

In the present configuration, speech is digitally recorded and saved on disk, as described above, using the Raytheon 704. The 704 then creates an A-matrix from the digitized waveform. The A-matrix is then sent (via direct hardware link) to the IBM 370/145, which then performs all subsequent linguistic processing of the utterance and returns a response to the user. Since the system is dependent upon thematic patterning for assistance in understanding an utterance, it is necessary for the user to interact with VDMS using goal-directed dialogs. For testing purposes, ten dialogs were created, with an average of ten utterances per dialog. Each of two male speakers (for whom vowel formant tables had previously been generated) recited the sets of dialogs. In this initial test of VDMS, an average of 52% of the utterances were correctly understood. Analysis of these preliminary results has shown that this figure can be increased by implementing some modifications to the phonological processes and lexical matching procedure.

2.2.7 Related Research

A continuing program of basic acoustic-phonetic research is providing algorithms that will improve the over-all accuracy of current and future speech-understanding systems. An experiment designed to compare the F1 and F2 frequency movements of vowels next to /r/ with the same vowels before other consonants was conducted (Kameny, 1974). Lehiste's (1964) data (obtained from spectrograms) on the vowel allophones associated with /r/ were used for comparison purposes. The data for this experiment were based on formant trajectories computed by LPC techniques on the Raytheon 704. The results of this experiment confirmed Lehiste's work, which indicates that there is a change in some vowels in a retroflexed environment. The change in vowels after /r/ is minimal except for /i/, but the change in vowels before /r/ is considerable. This was a preliminary experiment in which the number of subjects and samples was small. However, the results can be used to develop a retroflexed vowel space on the basis of a non-retroflexed vowel space and to compare the identification of vowels using this new F1-F2 space with the identification of vowels using the non-retroflexed F1-F2 space.

An algorithm that automatically distinguishes the nasals /n/, /m/, and /ŋ/ from each other was designed (Gillmann, 1974; Gillmann and Ritea, 1974). Spectral analysis is performed on the Raytheon 704 using an LPC model to locate the formants of these phonemes. By comparing the formant frequencies of unknown nasals to prototype values derived from normalization utterances, the algorithm was able to correctly identify nasals in 72% of the cases tested. Experimentation has indicated that (1) automatic techniques can be employed to distinguish nasals in continuous speech; (2) linear prediction can be used effectively to analyze the spectra of these phonemes; and (3) speaker-dependent tables of prototype nasal formants extend these results to multiple-speaker environments.

2.3 PROGRESS FOR THE FINAL SIX-MONTH PERIOD

The original goal to be reached by the end of the current contract year (as specified in SDC Proposal 73-5674) consisted of enlarging the vocabulary of the query language and loosening the grammar to make the language easier and more natural to use. Specifically, VDMS was to contain a vocabulary of about 500-600 words, and the grammar was to be modified to admit the following capabilities:

(1) A more natural form of expression for integers (for example, "thirty four" instead of "three four").

(2) A facility for inter-item comparisons (for example, "Print category where surface speed greater than submerged speed.").

(3) Use of strings of inequalities (for example, "Print type where draft greater than seven and less than nine.").

(4) Simple arithmetic calculations (for example, "Print category where surface speed greater than <u>three times</u> submerged speed.").

(5) Interrogative sentences.

However, recent discussions and negotiations (at the request of ARPA) with the Stanford Research Institute (SRI) have yielded a cooperative research plan for further development of a joint speech understanding system. Within this plan, SDC is concentrating primarily on signal processing, acoustics, phonetics, phonology, and system software and hardware support. SRI is concentrating on syntax, semantics, pragmatics, and discourse analysis. System design and architecture, and prosodics, are shared concerns.

The first major goal to be achieved under this cooperative research plan will be the successful development, implementation, and demonstration of a speech understanding system in late 1974.

2.3.1 <u>Demonstration System Overview</u>

For the initial implementation of the 1974 demonstration system, which operates on the Raytheon 704 and IBM 370/145 computers, the task domain is data management, and the data base consists of information on the submarines fleets of the United States, the United Kingdom, and the Soviet Union. The acoustic-phonetic processing and lexical mapping routines are essentially as used in VDMS, modified to handle a vocabulary of 300 words. There is a word-string mapping procedure that handles coarticulation between pairs of words. The parser is a major revision and extension of the previous SRI parser (Paxton, 1974), in which sources of knowledge are separated from the procedures for applying them. A best-first strategy still prevails, but it is now possible to start from any fixed point in the utterance, to skip over portions, and to accept input from word-spotting routines. The grammar encompasses that of the previous SRI system but has been extended to cover isolated noun prases and nominals. In addition to being independent of the parser, it has been rewritten as a series of context-free rules with factors that specify restrictions or conditions on rule application; as a result, it can be used top-down, bottom-up, or with missing segments. The semantics have been completely revised from the previous SRI system. Now, information is stored in a network representation: corresponding to each syntactic rule there is a semantic interpretation rule that operates on the network. A pragmatic component, based on an analysis of protocol studies, has been added to handle anaphora and ellipsis and to provide discourse constraints for processing dialog dependencies. Although the acoustic-phonetic processor and the system hardware and software will undergo little or no change for this system, a fair amount of research has been accomplished in these areas.

2.3.2 Acoustic-Phonetic Research

For purposes of acoustic feature extraction, two programs were developed: one for automatic formant frequency analysis and another for fundamental frequency extraction.

The formant frequency analysis program assigns frequency, amplitude, and bandwidth to each of the first three formants for each 10-msec. voiced segment of continuous speech. Its input parameters are fundamental frequency, RMS, and up to five spectral peaks below 5,000 Hz. with their respective amplitudes and bandwidths. Peak information is obtained from linear prediction spectra. Techniques distinguishing this formant tracker from previously reported ones are that: (1) all spectrum computations are accomplished in the peak-picking phase, prior to formant tracking (this may yield spurious formants but few missing formants); (2) anchor points are located by selecting three consecutive 10-msec. segments in which three possible formant frequencies do not differ from one segment to the next by more than a threshold amount; and (3) decision-making is aided by frequency pattern matching when more than one formant is possible for a given slot (this is particularly useful for nasals and /l/ and /w/). Frequency-pattern information is derived from speaker vowel-sonorant frequency tables.

Fundamental frequency is extracted by a three-stage process. The speech is first digitally low-pass filtered and down-sampled from 20,000 samples per second to 2,000 per second. Autocorrelation spectra are then taken every 10 msec. using what Skinner (1973) describes as the "end-off" multiplication technique. Finally, a pitch-tracking pass extracts peaks from these spectra, refines them by parabolic curve fitting, and assembles these values into a coherent pitch track by editing out octave errors (mistaking a harmonic or subharmonic for the fundamental frequency) and isolated anomalies.

In addition to acoustic feature extraction, research was conducted on the acoustic correlates of style of speech. A major problem in speech research has been the selection of test material for both acoustic-phonetic experimentation and speech understanding system exercising. Some experimenters favor the use of read speech: others favor the use of spontaneously spoken speech. However, nothing had been done to determine the acoustic effects of these different styles of speech or to determine whether there were actual differences. Our goal was to construct a carefully designed experiment in which both read and spontaneous (as well as other speech styles) could be compared. To this end, recordings were made of ten speakers of "California English" producing the set of test words: "bee, bow, boy, bed, bad, bud" in seven different styles of speech, as follows:

(1) Free speech during an interview in which subjects were induced to say the test words without their having been previously spoken by the experimenter;

(2) Spontaneously spoken lists of words;

(3) Spontaneously produced sentences;

(4) Repetition of sentences spoken by the experimenter;

(5) Reading a continuous passage;

(6) Reading the test words in the sentence "The word is ___."; and

(7) Reading the test words in lists.

Each word was made to occur in phrase final, stressed position. The data were analyzed using the formant frequency analysis program to determine the frequency, amplitude, and intensity of the first four formants. The nucleus of each vowel in each word in each style was determined by an algorithm. The nucleus of the vowel differed in some styles of speech. Preliminary conclusions indicate that differences in the formant structure of vowels in read and spontaneous speech do exist.

2.3.3 System Hardware and Software

Acoustic feature extraction and phonetic processing for future SDC/SRI speech understanding systems will be done by linked PDP-11/40 and SPS-41 computers. (Experimentation to determine what processing is required will continue to be done on the Raytheon 704.) The PDP-11/40 was delivered in June. We are currently awaiting a completed version of the ELF operating system for the PDP-11/40, which is being prepared by SCRL. Once this is received, we will implement a number of user-level programs, which are now in a design stage. Delivery of the SPS-41 is currently scheduled for late November.

In late 1973, an ARPANET interface for the PDP-11 was developed at SDC. Designated the HSI-11A, this interface has been operational at SDC since January, 1974. In March, 1974, the designer (Lee Molho) was asked by the ARPA Interface Committee (ISC) to submit HSI-11A for possible selection as a standard ARPANET interface for PDP-11 computers. In May, the HSI-11A design was selected. For several months thereafter, ISC members and the SDC staff conducted technical discussions, primarily by ARPANET "Network Mail". The purpose of these discussions was to specify an HSI-11B design suitable for production by some organization for widespread, general use on the ARPANET. These discussions resulted in 11 engineering changes to the original HSI-11A design in order to meet ISC requirements.

In addition to system hardware and software efforts on the PDP-11/SPS-41 computer configuration, system support work for the SUR effort is being done in the form of the development of a programming language and system that is specifically designed for the implementation of SUR systems. This language, called CRISP, will be operational in July, 1975, on the IBM 370/145 under VM/370. A first draft of the language and system design document will be completed in December, 1974. CRISP offers a set of capabilities that, taken together, make it a uniquely appropriate tool for the implementation of our speech understanding systems. Among these capabilities are:

- A structured data capability similar to that available in PL/I.

- Flexible pointer manipulation similar to that in LISP, including functionals.

- Multi-processing and "spaghetti-stack" primitives.
- Efficient compilation of both arithmetic and pointer-manipulation algorithms (incremental and batch modes).
- Three levels of extensible languages available to the user:

  (1) Source Language (SL)--an ALGOL-like language with infix operators.

  (2) Internal Language (IL)--A LISP-like Polish-prefix list structure language.

  (3) Assembly Language (CAP)--a macro-assembly language.

- Availability of dynamic, local, and own variables.
- Name pooling.
- System aids to better utilize virtual memory resources.
- A variety of aids for group construction of large programs.

Also being prepared is a translator program that converts SDC Infix LISP to CRISP/SL.

Using CRISP, the bottom-end numerical algorithms, mapping procedures, and top-end component may all be combined in a single language and system without loss of efficiency. This is advantageous for several reasons, the most important being the increased ability of the modules to coordinate and communicate with one another.

2.4 PLANS

The objective for the 1974-1975 contract year is the successful operation of a milestone system, developed jointly by SDC and SRI, capable of handling utterances in ordinary English appropriate for task-oriented dialogues of the data management task. This milestone system will have the following characteristics:

- A vocabulary of approximately 600 words.
- Ability to accommodate six speakers (male and female).
- Response to the user in about 25 times real time.

- Operating-environment signal/noise ratio of approximately 30-40 dB.

- Accuracy in understanding at the utterance level of 90%.

For this joint system, SDC will assume primary responsibility for:

- Signal processing algorithms, which will perform acoustic feature extraction from the waveform on the PDP-11/40 and SPS-41 computers.

- Acoustic-phonetic analysis, which will include investigations of voiced fricatives and plosives, segmentation techniques, and development of procedures for assigning phonetic features to vowels, nasals, and sonorants.

- Lexical matching procedures, which will consist of the development of bottom-driving techniques, lexical subsetting methods to quickly prune unlikely candidates from lists of proposed words, and prosodic mapping techniques to map phrases using prosodic information.

- System hardware and software, including interconnection of the PDP-11/40 and SPS-41 computers, associated software development, and the creation of formal test and validation procedures for modules of the speech understanding system.

SDC and SRI will share responsibility for the system architecture and the analysis of prosodic information. SRI will have primary responsibility for:

- Protocols and discourse analysis.

- Parser development.

- Grammar and semantics.

The activities for which SDC has primary responsibility are described in detail in SDC Proposal 74-5490.

The objective for the 1975-1976 contract year is the successful demonstration of the Five-Year System (Newell, et al., 1971). For this system, a second task domain will be added. Each domain will have a vocabulary of 1,000 words. The system will accommodate about 30 speakers and will run on the PDP-11/40, SPS-41, and IBM 370/145 computers.

2.5 STAFF

H. Barry Ritea, Project Leader

Jeffrey A. Barnett
William A. Brackenridge
Richard A. Gillmann
Iris Kameny
Peter Ladefoged (Consultant)
Lee M. Molho
Douglas L. Pintar
Georgette Silva
Rollin V. Weeks

3. LEXICAL DATA ARCHIVE

3.1 INTRODUCTION

The Lexical Data Archive (LDA) project has addressed itself to the task of providing the ARPA Speech Understanding Research (SUR) projects with semantic and syntactic data for the words in their lexicons. Being devoted exclusively to lexical research, the LDA project can assure a broad range of services for each SUR project without the triplication of effort and resources that would be required if the three SUR projects were to collect and analyze these data themselves. The project is monitoring a broad range of lexical data sources, selecting the data having potential payoff for speech understanding, formatting those data for archival purposes, and providing for their dissemination to the appropriate SUR projects. The data in the archive are centered on the 3,000 or so words appearing in the twelve lexicons currently being used by the SUR projects at Bolt Beranek and Newman Inc., Carnegie-Mellon University, and System Development Corporation. Although there is considerable overlap among the lexicons, the words treated come from quite disparate domains: chess playing, analyses of moon rocks, submarine fleet data, project management, and the daily news releases of the Associated Press.

3.2 PROGRESS AND PRESENT STATUS

Since the initiation of the LDA project in October, 1973, the following six tasks have been pursued:

(1) Completion of the design of the Semantically Oriented Lexical Archive (SOLAR).

(2) Collection of data from the linguistics and philosophical literature.

(3) Development of computer programs for deriving machine-readable data files.

(4) Development of computer programs for discovering and displaying links among words in particular lexicons.

(5) Development of a file-management facility.

(6) Distribution of data to the SUR projects.

These tasks are described in the following sections.

3.2.1 Completion of SOLAR Design

First, the design of the Semantically-Oriented Lexical Archive (SOLAR) has been completed. This includes the decision as to the types of lexical data to be collected, the determination of the data collection procedures, the specification of programs needed to extract data from machine-readable transcripts, and the design of the logical structure of the files to be built. In accordance with the responses from the SUR groups to a questionnaire, initially distributed in June, 1973, SOLAR will consist of ten files (seven of which have been implemented wholly or in part):

(1) A word index, which allows a user to easily determine the words for which data are being collected and the types of data currently available for a given word.

(2) A bibliographic reference file, which can be used as a resource for accessing the literature and which can also be used in conjunction with other files to abbreviate references within SOLAR.

(3) A file of semantic analyses, which contains formal treatments of the semantic properties of individual words as found in the literature. This file is being built manually, by locating and reading documents relevant to the SUR lexicon words, extracting the essence of each document's analysis, writing a critique of it, and entering this information on sheets for keypunching. Although each analysis of a particular word is treated separately, the analyses are tied together by cross-referencing in the critiques appended to the analyses.

(4) A file summarizing the theoretical backgrounds from which the semantic analyses have been extracted.

(5) A file explaining and commenting on the descriptive constants employed in the semantic analyses.

(6) A file of integrated summaries of analyses given in the literatures of philosophy and artificial intelligence for concepts invoked by the descriptive constants.

(7) A file of collocational information found in the definitions of Webster's Seventh New Collegiate Dictionary (W7), which has been machine-extracted and is accessible via the words to which it pertains.

(8) A file of definitional links between words within a particular SUR lexicon. These are being constructed so

that a user can observe the semantic interrelations in his lexicon.

(9) A file of semantic fields, which are being designed for each SUR word by tying to it words found in certain definitional, synonymitive, and antonymitive relationships in W7, _Webster's New Dictionary of Synonyms_ (WNDS), and _Roget's International Thesaurus_ (Roget).

(10) Finally, every context of each SUR word as found in the W7 definitions, in the Brown Corpus, and in selected speech dialogs is being entered in a keyword-in-context (KWIC) file.

For a more detailed discussion of the contents of each of these files, see Diller and Olney (1973).

3.2.2 _Hand Collection of Data_

A significant amount of data has been collected for the hand-built files (files 1-6, above). The approximately 3,000 SUR words existing as of September, 1974, have been entered into the word index together with their W7 parts of speech and an indication of the SOLAR data available for each. About 2,700 references to documents in the linguistics and philosophical literature have been collected by LDA personnel and entered into the bibliographic file. Slightly more than 1,600 other citations to articles in experimental phonetics, psychoacoustics, speech analysis and synthesis, and phonology have been entered through a data exchange with Bell Laboratories. Another 1,000 entries in psycholinguistics and phonology have been entered through an arrangement with the UCLA Department of Linguistics. Indexing by author, title, and keyword (among other parameters) is possible for each bibliographic entry. Approximately 300 semantic analyses have been written, and about 150 have been converted into machine-readable data sets to facilitate updating and distribution. (Section 3.2.2, titled "Sample Semantic Analyses", of the interim report (Bernstein, 1974) presents four such semantic analyses.) Explanations of about 200 descriptive constants used in the semantic analyses have been written, and about 100 have been computerized. (In Section 3.2.3 of the interim report, some sample explanatory notes were given.) The file containing integrative summaries of conceptual analyses has received considerable attention in recent months. Approximately 20 summaries have been written; half of them have been computerized and are now being translated into a formal-logic language to facilitate their incorporation into the knowledge structures of the SUR systems. (In Section 3.2.4 of the interim report, three sample conceptual analyses were presented.)

3.2.3 Machine Derivation of Data

Several programs have been developed to allow the creation of two of the four machine-derived data sets. First, a set of programs was written that restructured the W7 parsed transcripts into a format suitable as input to other programs. This set includes one program that reassembles the definitions from their parsed format and another that extracts from all of W7 just that subset of definitions relevant to the current list of SUR words. Second, a program for building the collocational feature file was written, compiled, debugged, and run. The resulting file contains about 10,000 lines of text containing W7 definitions that show permissible contextual features for particular senses of the SUR words. Third, the program that converted the Brown Corpus KWIC data set to SOLAR format was written and run. Since we limited the number of sample contexts per word to a maximum of 450, the resulting file contains about 300,000 lines. The addition of the W7 definitional contexts is awaiting an evaluation of the utility of the Brown contexts. We are currently adding contexts from dialogs collected by the SUR groups.

3.2.4 Programs under Development

Considerable progress was made in creating the programs and data sets needed to build the file displaying definitional links between words in particular lexicons. The data sets being built comprise the words particular to a given lexicon, the syntactic parts of speech for each, the W7 definitions for each, words standing in an inflectional relationship to the core lexicon, and a list of stop words for which no definitional links are followed.

Work on the file of semantic fields has centered mainly on the definition of a data structure and the collection of relevant data sets. Keypunching of the antonym relations found in WNDS began in October, 1974. The quasi-synonymitive relations found in Roget must await the release of the Roget transcript by the Sedelow group at the University of Kansas. Prof. Sedelow expects to complete the editing of the transcript near the end of 1974.

3.2.5 Data Management

The six files that are currently computerized are accessible via CDMS, an SDC data management system with exceptional update and report generation capabilities. This system has greatly facilitated the creation of the hand-derived files. However, since the system will not be available after October 31, 1974, considerable effort was spent late in contract year 1974 in preparing to move all SOLAR data to another SDC data management system that is accessible via the ARPANET. The logical structure of all SOLAR files was reevaluated and revised where necessary, and the first of eight programs needed to convert the data sets to the revised format was coded and run.

3.2.6 Data Distribution

Early this year, the archive was publicized throughout the United States, Canada, Australia, and Europe. Approximately 20 researchers responded to our solicitation for documents dealing with lexical semantics, and about 35 expressed interest in receiving data from the archive.

In April, 1974, initial distribution of author and keyword indices to the bibliographic citation file was made to the SUR projects and to five university linguistics departments. In October, 1974, a revised listing of the citation file was distributed, together with initial listings of the word index, the semantic analyses, the descriptive constants, the conceptual analyses, the collocational features, and portions of the KWIC file.

3.3 PLANS

During the next (1974-75) contract year, the LDA staff will focus on five tasks. First, we will continue data collection from the literature. This will involve extending the bibliographic files, more than doubling the semantic and conceptual analysis files, and updating the word index. The updating activity derives from the continual addition of words to the SUR lexicons.

Second, we will continue program development. Some restructuring of files and programs is expected as a result of feedback from users regarding the utility of each file. We will also be coding and running the programs needed to produce the two remaining machine-derived files (i.e., the file linking words definitionally and the file of semantic fields). Included in this effort will be the keypunching of antonymitive relations found in WNDS. These will be added to the semantic field file to permit their incorporation into the semantic networks of the SUR systems.

Third, we will complete the work necessary to put SOLAR on the ARPA Network and distribute user's guides indicating on-line accessing procedures. Fourth, we will continue to disseminate data from each of the files. Lastly, to facilitate use of the archive, we will continue to document the archive (producing further user's guides) and will provide demonstrations of the use of SOLAR.

In the succeeding (1975-76) contract year, the LDA staff will focus on the following tasks:

(1) Refine the data management output to tailor it more directly to the specific SUR projects requesting data (i.e., improve the selectiveness of data dissemination).

(2) Revise the semantic field file displays (and perhaps the file structure) in accordance with suggestions from users as to how the utility of the file could be enhanced.

(3) Update each of the files on the basis of the new words added to the SUR lexicons.

(4) Continue hand collection of data for the bibliographic reference, semantic field, and conceptual analysis files.

(5) Explore the limits of algorithmically producing a semantic network for a given lexicon from the data residing in SOLAR.

3.4 STAFF

Dr. Timothy C. Diller, Project Leader

John Olney (Consultant)
Thomas Bye (part-time)
Frank Heath (part-time)
Martin Mould (part-time)
Nathan Ucuzoglu (part-time)

4. COMMON INFORMATION STRUCTURES

4.1 INTRODUCTION

The Common Information Structures project is addressing the problem of converting and transferring data bases among disparate data management systems (DMSs). The needs for sharing data for different applications and for transferring existing data into new computer systems make it apparent that general techniques for data base conversion are desirable. It is equally desirable that these techniques be relatively easy to implement and use. The goal of this project is to develop techniques for data base conversion that are practical for applicaton to current data management systems and that are designed to be used easily by data base users.

The difficulties in converting a data base from one data management system (DMS) to another arise from the fact that data base structures are system and application dependent. As a result, a DMS imposes constraints on the form of the data bases residing in it. These constraints are of three types: (1) logical-level constraints, such as level of hierarchies, size and number of fields, and data types; (2) storage-level constraints, such as inversion and access paths of files and file indexing organizations; and (3) physical-level constraints, such as physical devices used and block/record structures. These levels were described in reports issued for the 1972-73 contract year.

The conventional method of converting data bases for new applications is to write a special-purpose conversion program for each data base. Another possible approach is to define data description languages for all three levels (logical, storage, and physical), then specify in these languages the source and target data bases, as well as conversion statements between them. This approach has been discussed by several researchers. (See Smith, 1971; Ramirez, 1973; Stored Data Definition and Translation Task Group, 1972; Smith, 1972; Fry, et al., 1972.) Since this approach involves all three levels, it requires complex and detailed data description languages, which are difficult to learn and to use. It also requires that data be converted from the source physical environment to the corresponding target physical environment, which further complicates any possible implementation.

The approach being taken by this project is based on the assumption that the data conversion process can depend mainly on conversion at the logical level. Conversion at this level can be achieved by using existing query and generate capabilities of DMSs to move data from their physical representation to the logical level and vice versa. The tasks required in the data conversion process are diagrammed in Figure 4-1. First, the source data base is retrieved, via the query capabilities of the source DMS, and reformatted into a standard form. Then, the data translation process takes place, and a target data base in the standard form is produced and reformatted into a data format acceptable to the generate capability of the target DMS. Finally, the target data base is generated with the generate capability of the target DMS.

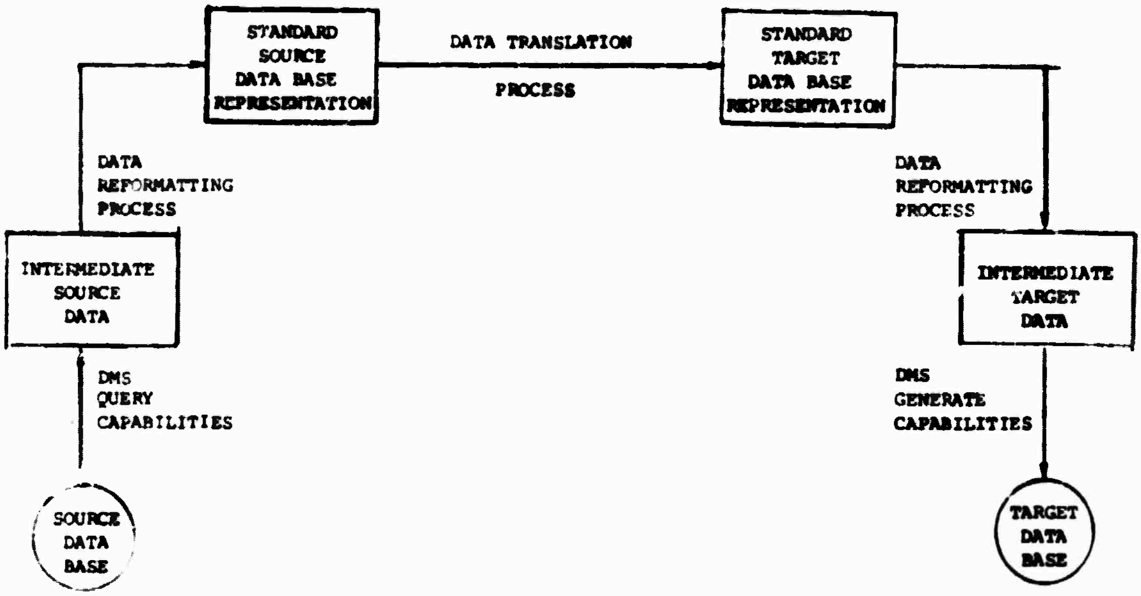
STANDARD SOURCE DATA BASE REPRESENTATION
DATA TRANSLATION PROCESS
STANDARD TARGET DATA BASE REPRESENTATION
DATA REFORMATTING PROCESS
DATA REFORMATTING PROCESS
INTERMEDIATE SOURCE DATA
INTERMEDIATE TARGET DATA
DMS QUERY CAPABILITIES
DMS GENERATE CAPABILITIES
SOURCE DATA BASE
TARGET DATA BASE

Figure 4-1. The Data Conversion Process

4.2 PROGRESS AND PRESENT STATUS

The first task of this project during the present contract year was to explore the possibilities of automatically generating target data descriptions from source data descriptions. We studied the file definition languages (FDLs) of several DMSs in an attempt to isolate the restrictions these systems impose on data structures. After thoroughly analyzing these FDLs, we concluded that automatic generation of a target description from the source description is either:

(a) Trivial--when the restrictions on the complexity of data base structures of the target system are less constraining than the restrictions of the source system (for example, going from a system that allows one level of hierarchy into a system that allows nine levels of hierarchy): or

(b) Impossible to predict in the general case, because the target description is semantically dependent on the intended use of the data base and on considerations of time, space, and cost.

Consequently, we concluded that it would not be fruitful to continue in this direction, and we decided to direct the main effort of the project at the development of processes that actually convert data, rather than at the automatic generation of target data descriptions. In the context of developing the details of the data conversion process, we performed the several tasks described in the following paragraphs.

A detailed study was made of the different data description languages and data conversion approaches described in the literature. (See Sibley and Taylor, 1973; Taylor, 1971; Smith, 1971; CODASYL Systems Committee, 1971.) We concluded that because these approaches advocate the use of details of all three levels of data description, they are too difficult to implement and use. This led us to the current approach, which depends mainly on the logical level of data.

A methodology for the data conversion process was developed. It involves the development of source and target reformatters and a logical data translator. These processes are driven by statements written in three languages, which are dependent mainly on logical characteristics of the data to be converted. These languages are the Common Data Description Language (CDDL), the Common Data Translation Language (CDTL), and the Common Data Format Language (CDFL).

A preliminary version of CDDL was developed. This was accomplished by analyzing existing DDLs (especially those of CODASYL and Smith) and selecting the subset that is relevant to our approach. The syntax of CDDL represents a logical view of hierarchical data structures. Because it contains no storage-level or physical-level requirements, it is a simple language for a user to specify. a file statement consists of group and field statements. A group statement, in turn, consists of field statements and, possibly, additional group statements, thus establishing a hierarchical structure. There are several types for fields, and a repetition number for groups puts an upper limit on the number of values in a group instance. The statements in CDDL, together with statements in CDTL, supply all the information necessary for the data conversion process. As a consequence, CDDL might change according to the requirements of CDTL.

The functions necessary for CDTL were developed. These functions are represented in terms of conversion statements between fields of the source data description and the target data description. A conversion statement consists of an association of one source field (or more) to a target field, plus an algorithm for the conversion of data (such as truncation, concatenation, or data-type transformation). Types of conversion statements were identified and form the basis of CDTL. In brief, these types are as follows:

(1) Instance -- represents a mapping of instances of a field or a repeating group (RG) into a field of a higher-level RG.

(2) Bundle -- represents a combination of multiple values of fields of the same RG level into an RG instance of a lower level.

(3) Operation -- allows a set of values in an RG instance to be combined by some operation (e.g., Average) into one value of a field in a higher-level RG.

(4) Direct -- allows for an association of source and target fields according to a given algorithm (e.g., truncation).

(5) Repeat -- necessary when a repetition of a field value through values of a lower-level RG is required.

(6) Levelup -- can be used to create an upper-level target RG from a source RG that has repeating values.

(7) Concatenation -- necessary when a target field is made up by concatenating source fields (or portions of them).

(8) As-is -- used when a portion of the source data is to be moved unchanged to the target data.

(9) Inversion -- necessary when an alternative view of the data base is required (e.g., a department-employee data based needs to be reorganized as an employee-department data base).

The identification of conversion types led to the problem of determining which mapping types can meaningfully coexist. We discovered properties that guide us in identifying those combinations that are semantically and logically sound and those that should be rejected. For example, if there is a DIRECT between a field of a source RG and a field of the corresponding target RG, then no other type between fields of those RGs is possible. We defined the concept of "correspondence" between source and target RGs; using it, we could talk about "up" mappings and "down" mappings. Thus, combinations of "up" and "down" mappings cannot coexist between fields of the same source and target RGs. The outcome of this stage of semantic analysis was the definition of the CDTL, which expresses conversion mappings in terms of field-to-field mappings only. This greatly simplifies specifying the required conversion mappings.

A version of the CDTL was defined. For reasons explained above, it is a fairly simple language, consisting mainly of field-to-field mappings between source and target data structures; this simplicity is a most important property for users. Another feature that we included in the CDTL is the ability to specify a string modification. This facility allows field values to be modified and reconfigured in a way similar to that of the Data Reconfiguration Service (Anderson, et al., 1971). However, we did not find it necessary to include facilities such as explicit conditionals.

The design of the standard data format was completed. Two properties that we found important had to be compromised because of conflict: (1) an efficient way of reading values in a data record, using hierarchy levels and instance occurrence, and (2) the need to leave a portion of the data virtually unchanged when the AS-IS function is specified (i.e., a portion of the source data to be converted must be left unchanged, so that the converter can integrate this portion of the source data into the target data being formed). We discovered an elegant way of compromising these requirements, by defining a top-to-bottom, left-to-right linearization of a hierarchy instance, together with embedded relative displacements linking instances of the different levels. A detailed description of this structure will be presented in a future document.

The design of the translator was developed. This major task consists of two parts. The "front end" performs a lexical and semantic analysis of the CDDL and the CDTL statements, using the mapping properties referred to above. It should detect any logical or semantic inconsistencies and produce a "conversion table" for the second part, the "data converter". The conversion table contains step-by-step instruction entries that drive a controller. The controller then invokes the appropriate routines to perform the mapping functions. The data converter operates on a source data instance in the standard form and generates a target data instance in the standard form, according to the CDTL specifications.

Most of the modules of the translator have been designed and are outlined in flowchart form. Using them, we plan to start implementation of the translator in PL/I in October, 1974.

4.3 PLANS

Our goal for the next contract year is to implement a prototype of the data base translator, as well as to develop and design the source and target data reformatters. The following tasks are planned.

4.3.1 Implementation of the Logical Data Translator

This task is broken into two subtasks:

(a) Lexical and semantic analysis of the source and target data base descriptions in CDDL and the translation statements in CDTL. This step will produce internal tables that represent the data translation operations to be performed. The analyzer will operate in the following manner: After the CDDL and CDTL statements are read and checked for their syntax legality, the semantic analyzer checks whether the translation requests are semantically possible; the translation tables are then produced.

(b) Translation of the source data (in their standard form) to the target data (in standard form). In this step, the translator uses the translation tables produced in the previous step as follows: A controller reads table entries in sequence and interprets the field-to-field type mappings to be performed. It invokes an appropriate module (one for each of the mappings--DIRECT, INSTANCE, etc.), which in turn calls a "read" module to extract the appropriate data from the source records. After the desired value is obtained, the controller invokes a "write" module to generate the desired part of the target record. This operation repeats until the complete target record is generated. Records are generated until all of the source data have been translated.

Two additional modules are planned. One module operates before the translation process starts and is called the "subsetting" module. This operation is required when we want to consider only a subset of the data base for translation (e.g., only the female employees). The other module is for post-translation ordering and is used to order the target records after translation has been completed.

4.3.2 Design of the Reformatters

This task will require a study of the common input and output data formats of data management systems. We project that only a few basic formats will be identified. These formats could be used to define a Common Data Format Language (CDFL). The implementation of data reformatters would then use a data format specification in CDFL to perform the reformatting. Another alternative is to build special-purpose reformatters for every data format type, an alternative that seems to be the more attractive choice when the number of format types is small.

4.3.3 Experimental Conversion of a Data Base

Toward the end of the contract year, we plan to experiment with the data translator by taking a small but useful data base from ARPA-DMS and converting it into a form acceptable to the Datacomputer. Initial discussions with the people involved at ARPA and CCA have been held.

4.4 STAFF

Dr. Arie Shoshani, Project Leader

Kenneth J. Fogt

5. REFERENCES

Anderson, R. D., et al. 1971. The data reconfiguration service -- an experiment in adaptable, process to process communication. ACM/IEEE Second Symp. on Problems in the Optimization of Data Communication Systems. New York: Association for Computing Machinery.

Barnett, J. A. 1974a. A phonological rule compiler. IEEE Symp. Speech Recognition: Contributed Papers. Pp. 188-92. New York: Institute of Electrical and Electronics Engineers, Inc.

------. 1974b. Module linkage and communication in large systems. IEEE Symp. Speech Recognition: Invited Papers. New York: Institute of Electrical and Electronics Engineers, Inc. In press.

Bernstein, M. I. 1974. Interactive systems research: interim report to the Director, Advanced Research Projects Agency, for the period 16 September 1973 to 15 March 1973. Report No. TM-5243/001/00. Santa Monica: System Development Corp.

CODASYL Systems Committee. 1971. Feature analysis of generalized data base management systems. New York: Association for Computing Machinery.

Diller, T. 1974a. SOLAR bibliography user's guide. Report No. TM-5292/000/01. Santa Monica: System Development Corp.

------. 1974b. The role of lexical semantics in automated speech understanding. In preparation.

Diller, T., and Olney, J. 1973. SOLAR: a Semantically Oriented Lexical ARchive. Report No. SP-3726. Santa Monica: System Development Corp.

------. 1974. SOLAR (a semantically oriented lexical archive): current status and plans. Computers and the Humanities. In press.

Fry, J. P.; Frank, R. L.; Hershey, E. A., III. 1972. A developmental model for data translation. Proc. 1972 ACM SIGFIDET Workshop. Pp. 77-106. New York: Association for Computing Machinery.

Gillmann, R. A. 1974. Automatic recognition of nasal phonemes. IEEE Symp. Speech Recognition: Contributed Papers. Pp. 74-79. New York: Institute for Electrical and Electronics Engineers, Inc.

Gillmann, R. A., and Ritea, H. B. 1974. Automatic isolation and analysis of nasals and nasalized vowels in continuous speech. *J. Acoust. Soc. Amer.* 55(Supplement):S21 (abstract).

Kameny, I. 1974. Comparison of the formant spaces of retroflexed and non-retroflexed vowels. *IEEE Symp. Speech Recognition: Contributed Papers.* Pp. 80-84. New York: Institute for Electrical and Electronics Engineers, Inc.

Kameny, I., and Weeks, R. 1974. An experiment in automatic isolation and identification of vowels in continuous speech. *J. Acoust. Soc. Amer.* 55(2):M12 (abstract).

Lehiste, I. 1964. *Acoustical characteristics of selected English consonants.* The Hague: Mouton & Co.

Molho, L. M. 1974. Automatic recognition of fricative and plosives in continuous speech using a linear prediction method. *J. Acoust. Soc. Amer.* 55(2):M13 (abstract).

------. 1974. Automatic recognition of fricatives and plosives in continuous speech. *IEEE Symp. Speech Recognition: Contributed Papers.* Pp. 68-73. New York: Institute of Electrical and Electronics Engineers, Inc.

Newell, A.; Barnett, J.; Forgie, J.; Green, C.; Klatt, D.; Licklider, J. C. R.; Munson, J.; Reddy, R.; Woods, W. 1971. *Speech-understanding systems: final report of a study group.* Pittsburgh: Carnegie-Mellon University, Computer Science Department.

Paxton, W. H. 1974. A best-first parser. *IEEE Symp. Speech Recognition: Contributed Papers.* Pp. 218-25. New York: Institute of Electrical and Electronics Engineers, Inc.

Ramirez, J. A. 1973. Automatic generation of data conversion programs using a data description language. Ph.D. dissertation, University of Pennsylvania.

Ritea, H. B. 1974a. A voice-controlled data management system. *IEEE Symp. Speech Recognition: Contributed Papers.* Pp. 28-31. New York: Institute of Electrical and Electronics Engineers, Inc.

------. 1974b. Speech input to a data management system. *Proc. Speech Comm. Seminar (SCS-74), Stockholm.* Pp. 291-98. New York: John Wiley & Sons.

Sibley, E. H., and Taylor, R. W. 1973. A data definition and mapping language. *Comm. ACM* 16(12):750-59.

Skinner, T. E. 1973. Autocorrelation method for determining fundamental frequency. *UNIVAC Intercommunication* April 23.

Smith, D. P. 1971. An approach to data description and conversion. Ph.D. dissertation, University of Pennsylvania.

------. 1972. A method for data translation using the Stored Data Definition and Translation Task Group languages. *Proc. 1972 ACM SIGFIDET Workshop.* Pp. 107-24. New York: Association for Computing Machinery.

Stored Data Definition and Translation Task Group. 1972. An approach to stored data definition and translation. *Proc. ACM SIGFIDET Workshop.* Pp. 13-56. New York: Association for Computing Machinery.

Taylor, R. W. 1971. Generalized data base management system data structures and their mapping to physical storage. Ph.D. dissertation, University of Michigan.

Weeks, R. V. 1974a. Predictive syllable mapping in a continuous speech understanding system. *IEEE Symp. Speech Recognition: Contributed Papers.* Pp. 154-58. New York: Institute of Electrical and Electronics Engineers, Inc.

------. 1974b. Syllable-mapping strategy for a continuous speech understanding system. *J. Acoust. Soc. Amer.* 55(Supplement):S22 (abstract).

APPENDIX

The text of this report was prepared on-line using the CMS/EDIT facility provided under the IBM VM/370 system. This report is available in machine-readable form.